CSIRO PUBLISHING
www.publish.csiro.au/journals/fpb *Functional Plant Biology*, 2006, **33**, 789–797

Identification of candidate phosphorus stress induced genes in *Phaseolus vulgaris* through clustering analysis across several plant species

Michelle A. Graham[A], *Mario Ramírez*[B], *Oswaldo Valdés-López*[B], *Miguel Lara*[B], *Mesfin Tesfaye*[C], *Carroll P. Vance*[D,E] *and Georgina Hernandez*[B,E,F]

[A]USDA–ARS, Corn Insects and Crop Genetics Research Unit, Ames, IA 50010, USA.
[B]Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Ap. Postal 565-A Cuernavaca, Mor. México.
[C]Department of Plant Pathology, University of Minnesota, St Paul, MN 55108, USA.
[D]USDA–ARS, Plant Research Unit, St Paul, MN 55108, USA.
[E]Department of Agronomy and Plant Genetics, University of Minnesota, St Paul, MN 55108, USA.
[F]Corresponding author. Email: gina@ccg.unam.mx

This paper originates from a presentation at the Third International Conference on Legume Genomics and Genetics, Brisbane, Queensland, Australia, April 2006.

Abstract. Common bean (*Phaseolus vulgaris* L.) is the world's most important grain legume for direct human consumption. However, the soils in which common bean predominate are frequently limited by the availability of phosphorus (P). Improving bean yield and quality requires an understanding of the genes controlling P acquisition and use, ultimately utilising these genes for crop improvement. Here we report an *in silico* approach for the identification of genes involved in adaptation of *P. vulgaris* and other legumes to P-deficiency. Some 22 groups of genes from four legume species and *Arabidopsis thaliana*, encoding diverse functions, were identified as statistically over-represented in EST contigs from P-stressed tissues. By combining bioinformatics analysis with available micro / macroarray technologies and clustering results across five species, we identified 52 *P. vulgaris* candidate genes belonging to 19 categories as induced by P-stress response. Transport-related, stress (defence and regulation) signal transduction genes are abundantly represented. Manipulating these genes through traditional breeding methodologies and / or biotechnology approaches may allow us to improve crop P-nutrition.

Keywords: ESTs sequences, genomics, legumes, phosphate deficiency, stress.

Introduction

Building the foundations for common bean functional genomics

Common bean is the world's most important grain legume for direct human consumption. In Mexico, and other countries of Central and South America, beans are staple crops serving as the primary source of protein N in the diet (Broughton *et al*. 2003; Graham *et al*. 2003). In Latin America and Africa, the yield of bean production is low, in part because of disease and insect pressures but also because of edaphic constraints that include soil N and P deficiencies, soil acidity, and aluminum, manganese, and iron toxicities (Graham 1981; Graham *et al*. 2003). It has been suggested that 89% of soils in Latin America are deficient in N and 82% are deficient in P, with more than 500 million Ha of soil in this region with a pH of 4.5 or less (Sánchez and Cochrane 1980). Overcoming edaphic stresses and improving crop yield are high-priority goals. Identification of the plant genes involved in these processes will not only increase our knowledge of processes integral to crop productivity, but also will identify new targets for crop improvement.

Despite the importance of common bean as a crop legume, very little expressed sequence tag (EST) information is publicly available. In efforts to develop molecular tools and to provide an initial platform for comparative

Abbreviations used: EST, expressed sequence tag; P, phosphorus; TIGR, The Institute for Genomic Research.

 10.1071/FP06101 1445-4408/06/080789

functional genomics, we initiated a collaborative project on common bean ESTs sequencing. We previously reported the sequencing and assembly of 21 026 ESTs derived from two *P. vulgaris* genotypes — Mesoamerican Negro Jamapa 81 and Andean G19833. ESTs were derived from root nodules, roots from P-deficient plants, developing and mature pods, and leaves (Ramírez *et al*. 2005). Recently, Melotto *et al*. (2005) reported the sequencing of 5255 ESTs from *P. vulgaris* genotype SEL1308. These sequences were derived from two cDNA libraries developed from leaves and seedlings, inoculated or non-inoculated with the fungal pathogen *Colletotrichum lindemuthianum*. Here we report the *P. vulgaris* gene index elaborated after the combined contig (the consensus sequence for an assembly of ESTs corresponding to a given gene) analysis of all the common bean ESTs publicly available.

Identification of genes important in P acquisition

Recent advances in macro- and microarray technology have led to the identification of several genes involved in plant responses to P deficiency. High-density macroarray analysis was performed to evaluate gene expression in response to P-deficiency in white lupin (*Lupinus albus*, Uhde-Stone *et al*. 2003). White lupin adapts to P deficiency by the development of short, densely clustered lateral roots called proteoid roots. Nylon filter arrays with some 2000 ESTs from proteoid roots were performed to identify genes differentially expressed in P-deficient proteoid roots as compared to normal roots. Some 35–40 genes that are more highly expressed in −P cluster roots than in + P roots were identified, including genes involved in carbon and secondary metabolism, P scavenging and remobilisation, plant hormone metabolism, and signal transduction (Uhde-Stone *et al*. 2003).

More recently, a comprehensive survey of gene expression in response to P deprivation in *Arabidopsis thaliana* has been reported (Misson *et al*. 2005). For this study the whole-genome Affymetrix gene chip (ATH1) was used to quantify the spatio-temporal variations in transcript abundance of some 22 000 genes. Analysis of short-, medium-, and long-term P deprivation revealed a total of 866 differently expressed genes; 612 of these were induced. Genes involved in several biochemical pathways that are closely associated with plant responses to P deficiency were coordinately activated and repressed. The functional classification of the differentially expressed genes also included those involved in ion transport, signal transduction, transcriptional regulation, and growth and development processes (Misson *et al*. 2005).

Identification of genes important in phosphate stress across species

Datasets that identify ESTs responsive to P-deficiency have been developed in several species. In *Medicago truncatula*, three EST libraries (totaling 13 245 ESTs) are available from P-starved roots and leaves [The Institute for Genomic Research (TIGR), http://www.tigr.org; verified 6 July 2006]. In soybean (*Glycine max* and *G. sojae*), 5 429 ESTs are available from P-starved roots (TIGR). In *L. albus*, 3260 ESTs are available from different stages of P-starved proteoid roots (Uhde-Stone *et al*. 2003 and C Vance unpubl. data). In addition, we have generated 3165 ESTs from P-starved roots of *P. vulgaris* (Ramírez *et al*. 2005). In this report, we use statistical and cluster analyses of EST data to identify candidate genes potentially involved in P-starvation from *P. vulgaris* and other legumes. Through *in silico* analysis of ESTs from *P. vulgaris*, *M. truncatula*, soybean, *L. albus*, and *A. thaliana*, we have identified P-stress genes that are statistically over-represented. Genes identified from *P. vulgaris* will build a foundation for future research.

Material and methods

Processing and contig assembly of P. vulgaris *and* L. albus *ESTs*

To analyse the most complete *Phaseolus vulgaris* L. dataset possible, the ESTs sequenced by Ramírez *et al*. (2005) and Melotto *et al*. (2005) were considered together. EST sequences generated by Melotto *et al*. (2005) were downloaded from DbEST (http://www.ncbi.nlm.nih.gov/dbEST/; verified 6 July 2006). The sequences of both groups were assembled with a processing pipeline developed by the Center for Computational Genomics and Bioinformatics (CCGB) at the University of Minnesota, as described by Ramírez *et al*. (2005). The results of this analysis are shown in Tables 1 and 2.

Uhde-Stone *et al*. (2003) sequenced 2102 ESTs from 7- to 10-d-old and 12- to 14-d-old P-starved proteoid roots of *Lupinus albus* L. An additional 1140 ESTs have recently been generated from emerging proteoid roots (C Vance unpubl. data). Using the processing pipeline mentioned above, we assembled the *L. albus* ESTs into 409 contigs and and 1379 singletons (data not shown).

Identification of contigs statistically over-represented with ESTs from P-starved tissues

To identify genes important under conditions of phosphate stress, TIGR's soybean (GmGI version 12) and *Medicago truncatula* Gaertn. (MtGI version 8) gene indices were searched for EST libraries derived from phosphate-starved tissues (http://www.tigr.org/tdb/tgi/plant.shtml; verified 6 July 2006). Three libraries were identified from *M. truncatula*. MHRP- and rootphos(-), containing 2649 and 1953 ESTs respectively, were derived from P-starved roots; NF–PL, containing 8643 ESTs, was derived from P-starved leaves. In soybean, only a single phosphate-starved root library was identified (5429 ESTs). In addition to TIGR's gene indices, we also took advantage of the 3165 P-stressed root ESTs available from *P. vulgaris* (Ramírez *et al*. 2005).

Table 1. Contigging statistics of *Phaseolus vulgaris* ESTs

Tissues	ESTs in contigs >1	EST Singletons
MesoAmerican nodules	2537	1208
MesoAmerican pods	2043	904
MesoAmerican roots	1882	1283
MesoAmerican leaves	2072	605
Andean leaves (5′ and 3′)	2075	1168
Shoot (Sel 1308)	1667	802
Shoot (Sel 1308) inoculated with *Colletotrichum*	1515	813
Total ESTs	13 791	6787

Table 2. Characteristics of tissue-specific contigs from *Phaseolus vulgaris* ESTs

Tissue-specific contigs	Number of contigs >1	Average ESTs per contig	Average length	Largest contig (no. ESTs)
MesoAmerican nodule-specific	207	2.5	785.6	10
MesoAmerican pod-specific	87	3.9	748.3	64
MesoAmerican root-specific	190	2.5	736.6	11
MesoAmerican leaves-specific	29	2.5	751.4	8
Andean leaves-specific	121	2.9	814.9	26
Shoot (Sel 1308)	36	2.1	617.5	3
Shoot (Sel 1308) inoculated with *Colletotrichum*	40	2.2	575.5	7
Mixed tissue contigs	2173	5.5	897.3	269
All contigs	2883	4.8	861.3	269

Custom perl scripts were used to examine each contig / phosphate-starved library combination. For most species, each EST contig was examined only once, since only one P-starved EST library was available. For *M. truncatula*, each contig was examined three times, once for each of the three P-starved libraries. For each contig / library combination, a perl script was used to count four observed values: the number of P-starved ESTs from a particular library in and out of the contig and the number of 'other' ESTs in and out of the contig. 'Other' ESTs included all other EST libraries except those from other P-starved libraries, those whose tissue origin could not be determined, or those representing resequenced libraries.

Based on the four observed counts, a second perl script calculated the expected values based on EST frequency. If the observed and expected counts within a contig were greater than four, a chi-square association test (Dunn and Clark 2001) determined the statistical significance of the results. If any of the counts was less than four, statistical significance was calculated by the Fisher exact probability test (Siegel 1956). Each of these tests determined whether a contig has a true over-representation of P-starved ESTs or if the number observed could occur by chance. The Fisher exact test was used with counts less than four because it won't exaggerate probability estimates. If the probability obtained for a particular contig / library combination was less than or equal to 0.05, a contig / library combination was considered statistically over-represented with ESTs from the corresponding P-starved library. Using this approach a single contig could be statistically over-represented with P-starved ESTs from more than one library. An example of the analysis is shown in Fig. 1.

In the case of *L. albus*, all available ESTs came from P-starved proteoid root tissues. Since ESTs from other tissues were not available, statistical analysis of *L. albus* contigs could not be performed. However, these genes were included in our clustering analyses because they were derived from P-starved roots and many of the genes have been confirmed to show enhanced expression under P-starvation with macroarray analysis (Uhde-Stone *et al*. 2003; S Miller, C Vance unpubl. data). Results from *L. albus* macroarray and *Arabidopsis* microarray experiments (Misson *et al.* 2005) provide strong support for bioinformatics approaches.

Identification of A. thaliana *genes induced in leaves and roots of P-starved tissues*

Misson *et al*. (2005) used the *Arabidopsis thaliana* (L.) Heyhn. whole-genome Affymetrix gene chip (ATH1) to identify *A. thaliana* genes induced or repressed during P-starvation. In P-starved leaves, 404 genes were significantly (probability <0.05) up-regulated with at least 2-fold induction. In P-starved roots, 231 genes were significantly up-regulated with a minimum 2-fold induction of gene expression. In total, 494 unique genes were up-regulated during P-starvation in leaves and / or roots. The gene identifiers were used to download the corresponding sequences from The *Arabidopsis* Information Resource (TAIR, http://www.arabidopsis.org; verified 6 July 2006).

Single linkage clustering of homologous sequences across species

The sequences of statistically over-represented P-starved contigs from *P. vulgaris*, soybean, and *M. truncatula*, the *A. thaliana* P-starvation statistically up-regulated genes identified by Misson *et al*. (2005), and the sequences of the 409 contigs assembled from P-starved proteoid roots of *L. albus* ESTs were combined to give a dataset totalling 2097 sequences (Table 3). Genes important in P-stress across species were identified by a two-step approach. TBLASTX (Altschul *et al*. 1997) of the dataset against itself was used to identify homologous sequences (E-value cutoff of 10^{-4}). Single linkage clustering, described by Graham *et al*. (2004), assigned homologous sequences into groups. Groups that only contained sequences from *L. albus* were deleted as no statistical analyses had been performed on these sequences. Once group identifiers were assigned, all sequences in each group were compared to the Uniprot protein database (Apweiler *et al*. 2004) using BLASTX (Altschul *et al*. 1997) and an E-value cutoff of 10^{-4}. These results were used in functional annotation of the groups.

Results

Phaseolus vulgaris *gene index*

We previously reported sequencing five EST libraries from *P. vulgaris*. Since then, an additional 5255 ESTs have been deposited in DbEST (http://www.ncbi.nlm.nih.gov/dbEST/) by Melotto *et al*. (2005). The combination of these two datasets provides a single *P. vulgaris* gene index containing 20 578 ESTs. Of these, 6787 were classified as singletons and the remaining 13 791 assembled into 2883 contigs ranging in EST redundancy from 2 to 269 (Table 1) resulting in a 9670 unigene set. Library specific contigs ranged from 29 to 207 (Table 2). Data from this new build can be downloaded from our website (http://www.ccg.unam.mx/phaseolusest/; verified 6 July 2006).

Identification of contigs statistically over-represented with ESTs from phosphate-starved tissues

Using algorithms and statistical analysis we evaluated 31 928 contigs from soybean, 18 612 contigs from *M. truncatula*, and 2883 contigs from *P. vulgaris* for statistical over-representation of ESTs from P-starved libraries (Table 3).

The number of contigs that were over-represented under P-stress conditions ($P<0.05$) in the three species ranged from 247 to 543. In the case of *L. albus*, the 409 contigs used for cluster analysis came from P-stress-induced proteoid roots. These *L. albus* contigs have been used in macroarray analysis to identify genes induced during P-starvation. The 494 contigs up-regulated due to P-stress in *Arabidopsis* and used in our cluster analysis were identified by Misson *et al*. (2005).

A ***Phaseolus vulgaris* contig 3247**

Observed values	ESTs in contig	ESTs out of contig	Total ESTs
P-starved root ESTs	38.00	3150.00	3188.00
Other ESTs	4.00	17 595.00	17 599.00
Total ESTs	42.00	20 745.00	20 787.00
Expected values	**ESTs in contig**	**ESTs out of contig**	**Total ESTs**
P-starved root ESTs	6.44	3181.56	3188.00
Other ESTs	35.56	17 563.44	17 599.00
Total ESTs	42.00	20 745.00	20 787.00
χ^2 analysis	**χ^2 in contig**	**χ^2 out of contig**	**Total χ^2**
P-starved root ESTs	154.63	0.31	154.94
Other ESTs	28.01	0.06	28.07
Total χ^2	182.63	0.37	183.00
Degrees of freedom	1		
Probability	1.07E–41		

B **Representatives of Group 179**

Sequence name[A]	Statistically significant tissue	*P*-value
At4g12470[B]	P-starved leaves	5.00E–02
MtTC100494	P-starved leaves	9.98E–03
MtTC100581	P-starved roots	2.63E–02
MtTC106613	P-starved roots	2.59E–03
PvContig1804	P-starved roots	2.35E–02
PvContig2421	P-starved roots	5.52E–04
PvContig2964	P-starved roots	1.98E–06
PvContig3247	P-starved roots	1.07E–41

[A]The two letters in front of the name are species identifiers: At, *A. thaliana*; Mt, *M. truncatula*; Pv, *P. vulgaris*.

[B]Identified in microarray analyses by Mission *et al*. (2005). Significance $P<0.05$.

Clustering of P-starvation-induced genes across species

One of the negative aspects of the statistical approach described above is that the number of EST libraries available for a given species may impact the results. For example, in each gene, the analyses would examine the number of ESTs derived from P-starved tissues compared with the number of ESTs derived from other tissues. For example, in *M. truncatula* and soybean, ESTs from P-starved roots were compared to all other ESTs including untreated roots. Genes identified as statistically over-represented in P-starved roots are likely a response to P-starvation. In contrast, in *P. vulgaris* no ESTs were available from untreated roots. Therefore, the statistically over-represented sequences we identified may be the result of overexpression in roots in general and may not reflect a response to P starvation (see Fig. 1).

To aid in identifying P-starvation-induced genes from *P. vulgaris* and to identify genes conserved across species, we clustered ESTs statistically over-represented from *P. vulgaris* with those from *M. truncatula* and soybean. In addition, we included statistically significant P starvation-induced genes identified from *A. thaliana* microarrays (Misson *et al*. 2005) and ESTs developed from P-starved *L. albus* roots for which additional macroarray data are available (Uhde-Stone *et al*. 2003). We hypothesised that if a candidate gene identified from *P. vulgaris* was also induced in response to P-starvation in other species, it may be a high priority candidate for future research. Single-linkage clustering was used to assemble the 2097 P-starved sequences from *P. vulgaris*, *M. truncatula*, soybean, *A. thaliana*, and *L. albus* into 287 sequence-homology based groups. Groups ranged in size from 2 to 98 sequences and had representatives from 1–5 species. Groups of genes that were

Fig. 1. Identification of genes expressed in response to P-starvation across species. (*A*) In this example from *P. vulgaris* contig 3247, a chi-square association test (Dunn and Clark 2001) was used to determine whether P-starved ESTs were statistically over-represented within the contig. In the first step, observed values are reported for ESTs assembled in the contig and ESTs outside of the contig. In the second step, the frequency of P-starved ESTs overall was used to calculate expected values for ESTs assembled in and out of the contig. The chi-square test then measured if the observed values were significantly different from the expected values. (*B*) To further limit the number of candidate genes identified, single linkage clustering was used to identify sequences that were homologous to *P. vulgaris* contig 3247 and were also over-represented during P-starvation. While the sequences from *M. truncatula* and *P. vulgaris* were identified by the bioinformatic methods described above, the *A. thaliana* sequence was identified by Misson *et al*. (2005) as significantly up-regulated in response to P starvation from microarray experiments.

Table 3. The number of genes (contigs) from five plant species examined to derive those that are statistically over-represented in libraries from P-deficient tissues

Species	Genes / contigs examined[A]	Statistically over-represented genes / contigs[B]
P. vulgaris	2883	247
Soybean	31 928	543
M. truncatula	18 612	404
A. thaliana	494	494
L. albus	409	409
Sum	54 326	2097

[A]In order to identify genes involved in P-starvation, we used sequences from a variety of sources. For *Phaseolus vulgaris* and *Lupinus albus*, contigs from the most recent assembly were used for analysis (see Materials and methods). For soybean and *Medicago truncatula*, TIGRs GmGI (version 12) and MtGI (version 8) gene indices were used. For *Arabidopsis thaliana*, sequences identified by Misson *et al*. (2005) were used for analysis.
[B]For *Phaseolus vulgaris*, soybean, and *Medicago truncatula*, statistical analyses were performed by chi-square association (Dunn and Clark 2001) or Fisher exact tests (Siegel 1956). No statistical analyses were performed on *Lupinus albus* sequences. For *Arabidopsis thaliana*, Misson *et al.* (2005) identified genes statistically up-regulated during P-starvation (probability <0.05) with a 2-fold increase in expression in P-starved roots and / or leaves.

over-represented in four or five species are listed in Table 4. Genes found to be over-represented in four or five species from P-stressed libraries were considered as important candidates involved in adaptation to P-deficiency.

Can bioinformatic analyses be used to identify candidate genes involved in P stress from P. vulgaris*?*

The 22 groups of genes listed in Table 4 are over-represented in P-stressed conditions and are likely candidates to be involved in plant responses to P stress. Of the 22 groups, 20 include *A. thaliana* sequences identified by Misson *et al*. (2005) as significantly induced in response to P stress. In addition, eight of the groups are supported by macroarray data of *L. albus* P-stress-induced genes identified by Uhde-Stone *et al*. (2003). *Phaseolus vulgaris* sequences are found in 19 of the 22 groups. By combining our bioinformatic analyses with available micro / macro array technologies and clustering results across species, we identified 52 *P. vulgaris* sequences (represented in the 19 groups) (Table 5). Two of these *P. vulgaris* genes are induced in P-deficient roots from bean plants, as shown by RNA-blot analysis (Ramírez *et al*. 2005). The genes noted in Table 5 will be priority targets for future research.

Table 4. Single linkage clustering identifies homologous sequences across species that are important in response to phosphate starvation
Group numbers are assigned at random. The data in this table correspond to groups containing sequences from four or five species. Two-letter designators are used to describe the species found in each group. At refers to *Arabidopsis thaliana*, Mt refers to *Medicago truncatula*, Gm refers to soybean, La refers to *Lupinus albus*, and Pv refers to *Phaseolus vulgaris*. All Pv, Mt, and Gm sequences were identified as statistically over-represented in P-starved tissues. All *Arabidopsis thaliana* sequences were significantly up-regulated in response to P stress (Misson *et al.* 2005; microarray data). All *Lupinus albus* sequences designated by an asterisk (*) were identified as significantly up-regulated in response to P stress in macroarray experiments (Uhde-Stone *et al.* 2003). *Phaseolus vulgaris* sequences designated by two asterisks (**) were identified as induced in roots from P-starved plants, by RNA-blot analysis (Ramírez *et al*. 2005). Annotations were assigned by comparing all sequences in the group to the Uniprot database using TBLASTX and a cut-off of $E < 10^{-4}$

Group number	Species represented	Sequences in group	Group annotation
0	At Mt Gm La* Pv**	17	Aquaporin
2	At Mt Gm La Pv	6	Pectin methylesterase
11	At Mt Gm La* Pv	98	Protein kinase
17	At Mt Gm La Pv	30	Peroxidase
18	At Mt Gm La Pv	22	ABC transporter family
42	At Mt Gm La Pv	12	WRKY transcription factor
56	At Mt Gm La* Pv	23	Cytochrome P450
77	At Mt Gm La* Pv	14	Oxygenase
14	At Gm La Pv**	6	Protein phosphatase 2C
15	Mt Gm La Pv	9	RAB protein
21	At Mt Gm La*	11	Phosphate transporter
31	At Gm La Pv	10	MYB transcription factor
57	At Gm La Pv	5	Dihydroflavonol or cinnamoyl reductase
62	At Gm La Pv	4	Proline-rich extensin
63	Mt Gm La* Pv	7	Glyceraldehyde 3 phosphate dehydrogenase
64	At Mt Gm La*	21	Purple acid phosphatase
74	At Mt Gm Pv	7	4-Coumarate-CoA ligase-like
79	At Mt Gm Pv	6	Zinc finger protein
84	At Gm La Pv	10	Glycosyl hydrolase
179	At Mt La* Pv	14	Class 10 PR protein
257	At Mt La Pv	4	Cyclic nucleotide-binding transporter
286	At La Mt Pv	31	Chlorophyll *a* / *b* binding protein

Table 5. Candidate *Phaseolus vulgaris* genes identified as likely to be relevant for response to P starvation

Group number	Contig name	Contig length	Probability of over-representation in P-starved roots	Top Uniref100 TLASTX B	E-value
0	PvContig2458	997	5.52E–04	Q41975 Probable aquaporin TIP2.2 (*A. thaliana*)	1.00E–89
0	PvContig3195	1297	2.90E–06	Q506K1 Putative aquaporin (*P. vulgaris*)	1.00E–161
2	PvContig1817	797	2.35E–02	O04887 Pectinesterase-2 precursor (*C. sinensis*)	1.00E–111
11	PvContig545	946	2.35E–02	Q9C753 Serine / threonine kinase, putative (*A. thaliana*)	1.00E–148
11	PvContig918	447	2.35E–02	Q9SII6 Hypothetical protein At2 g17220 (*A. thaliana*)	6.00E–16
11	PvContig1048	750	2.35E–02	Q9MBG5 Similarity to calmodulin (*A. thaliana*)	5.00E–24
11	PvContig1180	797	2.35E–02	Q5XWQ1 Serine / threonine protein kinase-like (*S. tuberosum*)	2.00E–96
11	PvContig1785	658	2.35E–02	Q9M1Q2 Serine / threonine protein kinase-like protein (*A. thaliana*)	1.00E–61
11	PvContig1788	952	2.35E–02	Q5JCL0 Mitogen-activated protein kinase kinase MAPKK2 (*G. max*)	1.00E–135
11	PvContig1805	617	2.35E–02	O49840 Protein kinase (*A. thaliana*)	1.00E–45
11	PvContig2010	721	3.60E–03	Q9SCZ4 Receptor-protein kinase-like protein (*A. thaliana*)	5.00E–54
11	PvContig2749	1269	2.42E–03	O81390 Calcium-dependent protein kinase (*N. tabacum*)	1.00E–153
11	PvContig2764	776	2.83E–02	Q9AR93 Putative calmodulin-related protein (*M. sativa*)	4.00E–50
11	PvContig2825	1172	2.42E–03	Q8H0B4 Wound-induced protein kinase (*N. benthamiana*)	4.00E–99
11	PvContig2949	1350	1.35E–03	P43293 Probable serine / threonine-protein kinase NAK (*A. thaliana*)	1.00E–137
14	PvContig2577	988	5.52E–04	Q8SBC4 Protein phosphatase 2C (*A. thaliana*)	1.00E–100
15	PvContig1792	717	2.35E–02	Q9SXT5 Rab-type small GTP-binding protein (*C. arietinum*)	3.00E–98
17	PvContig1255	572	2.35E–02	O22443 Seed coat peroxidase precursor (*G. max*)	3.00E–32
17	PvContig1784	853	2.35E–02	O80822 Peroxidase 25 precursor (*A. thaliana*)	1.00E–111
17	PvContig1853	1060	3.60E–03	O23961 Peroxidase precursor (*G. max*)	1.00E–133
17	PvContig2404	807	3.60E–03	Q9XFL3 Peroxidase 1 (*P. vulgaris*)	8.00E–87
17	PvContig2938	1167	1.35E–03	O23961 Peroxidase precursor (*G. max*)	1.00E–154
18	PvContig1254	762	2.35E–02	Q93XA0 TGA-type basic leucine zipper protein TGA2.2 (*P. vulgaris*)	1.00E–122
18	PvContig1732	727	2.35E–02	Q9SJR6 Putative ABC transporter (*A. thaliana*)	1.00E–114
18	PvContig1808	573	2.35E–02	Q5W274 PDR-like ABC transporter (*N. tabacum*)	1.00E–63
18	PvContig1821	584	2.35E–02	Q9C6R7 ABC transporter, putative (*A. thaliana*)	5.00E–43
18	PvContig2406	1225	5.52E–04	Q9M9E1 Putative ABC transporter (*A. thaliana*)	1.00E–133

Table 5. (*continued*)

Group number	Contig name	Contig length	Probability of over-representation in P-starved roots	Top Uniref100 TLASTX B	E-value
31	PvContig1813	733	2.35E–02	Q4JL82 MYB transcription factor MYB48-2 (*A. thaliana*)	2.00E–20
31	PvContig1742	769	2.35E–02	Q4JL82 MYB transcription factor MYB48-2 (*A. thaliana*)	1.00E–07
42	PvContig655	821	2.35E–02	Q6R7N3 Putative WRKY transcription factor 30 (*Vitis aestivalis*)	7.00E–43
42	PvContig2572	849	1.28E–02	Q3LHK9 Double WRKY type transfactor (*S. tuberosum*)	3.00E–46
42	PvContig2941	1455	7.88E–05	Q2PJR9 WRKY78 (*G. max*)	1.00E–128
56	PvContig1790	645	2.35E–02	Q9XFX0 Cytochrome P450 monooxygenase (*C. arietinum*)	2.00E–76
56	PvContig2816	806	8.46E–05	Q8S4C0 Isoflavone synthase (*P. lobata*)	1.00E–111
56	PvContig3066	1671	7.15E–04	P93147 Cytochrome P450 81E1 (*G. echinata*)	1.00E–155
57	PvContig1789	999	2.35E–02	Q6L3K1 Putative cinnamoyl-CoA reductase (*S. demissum*)	1.00E–125
57	PvContig1801	912	2.35E–02	O65152 Putative cinnamyl alcohol dehydrogenase (*M. domestica*)	1.00E–115
62	PvContig3189	1151	3.05E–10	No BLASTX Hit	
63	PvContig3106	1268	3.24E–05	P34921 Glyceraldehyde-3-phosphate dehydrogenase (*D. caryophyllus*)	1.00E–156
74	PvContig1828	718	2.35E–02	Q8H8C8 Putative AMP-binding protein (*O. sativa*)	1 / E–92
77	PvContig1718	705	2.35E–02	Q9C938 Putative oxidoreductase (*A. thaliana*)	1.00E–56
77	PvContig1767	673	2.35E–02	Q84L58 1-aminocyclopropane-1-carboxylic acid oxidase (*C. arietinum*)	1.00E–113
77	PvContig2400	846	3.60E–03	Q9C939 Putative oxidoreductase (*A. thaliana*)	1.00E–59
77	PvContig2407	914	5.52E–04	Q9C939 Putative oxidoreductase (*A. thaliana*)	3.00E–64
79	PvContig1737	929	2.35E–02	Q6L4C8 Putative zinc finger protein (*O. sativa*)	6.00E–31
84	PvContig1835	589	2.35E–02	Q700B1 Non-cyanogenic β-glucosidase (*C. arietinum*)	1.00E–40
179	PvContig1804	690	2.35E–02	P25986 Pathogenesis-related protein 2 (*P. vulgaris*)	1.00E–81
179	PvContig2421	626	5.52E–04	P25986 Pathogenesis-related protein 2 (*P. vulgaris*)	5.00E–78
179	PvContig2964	727	1.98E–06	Q41125 Proline-rich 14-kDa protein (*P. vulgaris*)	9.00E–52
179	PvContig3247	795	1.07E–41	P25985 Pathogenesis-related protein 1 (*P. vulgaris*)	2.00E–82
257	PvContig2414	726	1.28E–02	Q8H6U3 Cyclic nucleotide-gated channel A (*P. vulgaris*)	1.00E–74
286	PvContig3084	695	1.08E–02	Q9FNV7 Auxin-repressed protein (*R. pseudoacacia*)	2.00E–34

Discussion

The main goal of our study was to identify candidate genes from *P. vulgaris* that may be important in adaptation to P-stress. Our analysis was initiated with 2883 contigs identified in P-stressed roots of *P. vulgaris*. Using statistical and cluster analyses of EST library composition, we identified 247 candidate contigs that were statistically over-represented in P-stressed roots. Given the fact that we only had data from P-stressed roots it was not possible to ascertain whether these contigs were statistically over-represented generally in roots or specifically in response to P starvation. Therefore, we postulated that analysis of contigs that are

statistically over-represented in other species, particularly legumes, may provide insight in to *P. vulgaris* genes that respond specifically to P-stress. Using this approach we have identified 52 potential P-stress candidate genes for future research (Table 5). These genes may have universal importance in plant adaptation to P-stress.

The genes over-represented in P starvation in four or five plant species (Table 4) belong to various functional categories. Experimental evidence supports the relevant role of some of the genes identified in plant physiological adaptation to cope with P starvation. For example, several phosphate transporter genes cloned and characterised nearly a decade ago are transcriptionally regulated depending on external P availability (reviewed by Raghothama 1999; Smith 2001). A comprehensive transcriptional analysis of P-stressed *Arabidopsis* (Misson *et al*. 2005) revealed that several genes, including members of the Pht1 family of P transporters, ATP-binding cassette (ABC) transporters, peroxidases, transcription factors, organic acid synthesis as well as genes involved in sulfolipid synthesis are induced during P starvation (Misson *et al*. 2005). In our study, several genes involved in P acquisition or transport and mobilisation, such as phosphate transporters, aquaporins, ABC transporters, and phosphatases, were identified as candidate genes for universal response to P stress (Tables 4, 5). Experimental evidence from both *L. albus* and *P. vulgaris* support these results. Macroarray analysis in *L. albus* showed that several transporter, organic acid synthesis, and purple acid phosphatase genes are induced in P-starved roots (Uhde-Stone *et al.* 2003). RNA-blot experiments have shown that aquaporin gene expression is induced in P-starved roots of *P. vulgaris* relative to nodules and roots from normal plants and phosphatase transcripts were only detected in P-starved roots (Ramírez *et al*. 2005).

Accumulation of active oxygen species resulting in oxidative stress is common to several abiotic stresses including deficiency of nutritional elements in several plant species (Bartoz 1997). After prolonged P starvation, *P. vulgaris* plants show several symptoms of oxidative stress such as increased lipid peroxidation and hydrogen peroxide concentrations, and higher catalase and peroxidase activities in P-deficient roots than control roots (Juszczuk *et al*. 2001). In agreement with these reports, we find that peroxidase genes are over-represented as P starvation contigs in *P. vulgaris*, as well as in the other four plant species (Tables 4, 5). In addition, Table 4 shows genes that are often induced in response to elicitors, microbial attack, or under abiotic stress, which may be relevant for plant responses to P starvation, such as pathogenesis-related (PR) proteins and cytochrome P450s.

Notably chlorophyll *a* / *b*-binding protein was over-represented in four of the five species analysed (Table 4). This observation would be consistent with the dark-green leaf coloration that frequently accompanies P stress (Reuter *et al.* 1997). Although overall growth is eventually reduced in P-stressed plants, new leaves are continually generated at the expense of older leaves. These newly generated leaves synthesise the light-capture apparatus; thus, chlorophyll *a* / *b*-binding genes remain highly expressed (Utriainen and Holopainen 2001).

Genes with possible function in regulation or signal transduction pathways, such as protein kinases, zinc finger proteins and transcription factors, are also over-represented in four or five of the datasets analysed (Tables 4, 5). Transcription factors and signal transduction genes that display enhanced expression during P-deprivation are likely to play important roles in other stress conditions. Our analysis, as well as that of Misson *et al*. (2005), identified both WRKY and MYB transcription factors. Rubio *et al*. (2001) noted that a conserved MYB TF is involved in P-starvation signalling both in plants (*Arabidopsis*) and algae (*Chlamydomonas*).

Soil P limitation is a primary effector of root architecture, which refers to the complexity of root spatial configurations that arise in response to soil conditions (López-Bucio *et al*. 2003). Elegant experiments with common bean coupled to simulation modelling have shown that phenotypic adaptations to P deficiency involve changes in root architecture that facilitate acquisition of P from the topsoil (Ge *et al*. 2000; Lynch and Brown 2001). Although changes in endogenous concentrations of growth hormones such as ethylene and auxins have been proposed to mediate modifications in root architecture (López-Bucio *et al*. 2003), the signal transduction pathways or regulatory cascades for this complex plant response remain unknown. Genes such as those reported here (Tables 4, 5) that belong to the functional category of regulation / signal transduction may be relevant for regulating universal plant responses to P deficiency, such as modification of root architecture. Interestingly, a NAK gene containing an miRNA binding site is among the genes noted in Table 5. Recently miRNAs have been implicated in the P-starvation response of *Arabidopsis* (Miura *et al*. 2005; Chiou *et al*. 2006).

Experiments are currently underway to confirm the relevant role of candidate genes for *P. vulgaris* (Tables 4, 5) in the response and adaptation to P starvation. These experiments will compare the transcript profile of roots and nodules from P-deficient bean plants with control plants using approaches such as macroarrays, RNA-blot analysis and real-time quantitative PCR. Preliminary studies of *P. vulgaris* P-response candidate genes by RT–PCR have shown a WRKY and peroxidase expression up-regulated by P-stress (M Ramirez, CP Vance unpubl. data).

Acknowledgments

This work was supported in part by USA Department of Agriculture, Agricultural Research Service CRIS 3640-21000-019-00D ‘Improved Nitrogen and Phosphorus

Acquisition and Use in Legumes,' and CRIS 3625-21220-003-00D, 'Functional and Structural Genetic Analysis of Soybean.' GH received a sabbatical fellowship from DGAPA-UNAM.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389–3402. doi: 10.1093/nar/25.17.3389

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, *et al*. (2004) UniProt: the universal protein knowledgebase. *Nucleic Acids Research* **32**, D115–D119. doi: 10.1093/nar/gkh131

Bartoz G (1997) Oxidative stress in plants. *Acta Physiologiae Plantarum* **19**, 47–64.

Broughton WJ, Hernandez G, Blair M, Beebe S, Gepts P, Vanderleyden J (2003) Beans (*Phaseolus* spp.) — model food legumes. *Plant and Soil* **252**, 55–128. doi: 10.1023/A:1024146710611

Chiou TJ, Aung K, Lin SI, Wu CC, Chiang SF, Su CL (2006) Regulation of phosphate homeostasis by MicroRNA in *Arabidopsis. The Plant Cell* **18**, 412–421. doi: 10.1105/tpc.105.038943

Dunn OJ, Clark VA (2001) 'Basic statistics: a primer for the biomedical sciences.' 3rd edn. (John Wiley & Sons: New York)

Ge Z, Rubio G, Lynch JP (2000) The importance of root gravitropism for inter-root competition and phosphorus acquisition efficiency: results from a geometric simulation model. *Plant and Soil* **218**, 159–171. doi: 10.1023/A:1014987710937

Graham MA, Silverstein KAT, Cannon SB, VandenBosch KA (2004) Computational identification and characterization of novel genes from legumes. *Plant Physiology* **135**, 1179–1197. doi: 10.1104/pp.104.037531

Graham PH (1981) Some problems in nodulation and symbiotic nitrogen fixation in *Phaseolus vulgaris*: a review. *Field Crops Research* **4**, 93–112. doi: 10.1016/0378-4290(81)90060-5

Graham PH, Rosas JC, Estevez de Jensen C, Peralta E, Tlusty B, Acosta-Gallegos J, Arraes Pereira PA (2003) Addressing edaphic constraints to bean production: the bean / cowpea CRSP project in perspective. *Field Crops Research* **82**, 179–192. doi: 10.1016/S0378-4290(03)00037-6

Juszczuk I, Malusa E, Rychter AM (2001) Oxidative stress during phosphate deficiency in roots of bean plants (*Phaseolus vulgaris* L.). *Journal of Plant Physiology* **158**, 1299–1305. doi: 10.1078/0176-1617-00541

López-Bucio J, Cruz-Ramírez A, Herrera-Estrella L (2003) The role of nutrient availability in regulating root architecture. *Current Opinion in Plant Biology* **6**, 280–287. doi: 10.1016/S1369-5266(03)00035-9

Lynch JP, Brown KM (2001) Topsoil foraging — an architectural adaptation of plants to low phosphorus availability. *Plant and Soil* **237**, 225–237. doi: 10.1023/A:1013324727040

Melotto M, Montenegro-Vitorello CB, Bruschi A, Camargo LEA (2005) Comparative bioinformatics analysis of genes expressed in common bean (*Phaseolus vulgaris* L.) seedlings. *Genome* **48**, 562–570. doi: 10.1139/g05-010

Misson J, Raghothama KG, Jain A, Jouhet J, Block MA, *et al.* (2005) A genome-wide transcriptional analysis using *Arabidopsis thaliana* Affymetrix gene chips determined plant responses to phosphate deprivation. *Proceedings of the National Academy of Sciences USA* **102**, 11 934–11 939. doi: 10.1073/pnas.0505266102

Miura K, Rus A, Sharkhuu A, Yokoi S, Karthikeyan AS, *et al*. (2005) The *Arabidopsis* SUMO E3 ligase SIZ1 controls phosphate deficiency responses. *Proceedings of the National Academy of Sciences USA* **102**, 7760–7765. doi: 10.1073/pnas.0500778102

Raghothama KG (1999) Phosphate acquisition. *Annual Review of Plant Physiology and Plant Molecular Biology* **50**, 665–693. doi: 10.1146/annurev.arplant.50.1.665

Ramírez M, Graham MA, Blanco-López L, Silvente S, Medrano-Soto A, Blair MW, Herrnández G, Vance CP, Lara M (2005) Sequencing analysis of common bean ESTs. Building a foundation for functional genomics. *Plant Physiology* **137**, 1211–1227. doi: 10.1104/pp.104.054999

Reuter DJ, Elliott DE, Reddy GD, Abbott RJ (1997) Phosphorus nutrition of spring wheat (*Triticum aestivum* L.). Effects of phosphorus supply on plant symptoms, yield, components of yield and plant phosphorus uptake. *Australian Journal of Agricultural Research* **48**, 855–868. doi: 10.1071/A96159

Rubio V, Linhares F, Solano R, Martín AC, Iglesias J, Leyva A, Paz-Ares J (2001) A conserved MYB transcription factor involved in phosphate starvation signaling both in vascular plants and in unicellular algae. *Genes & Development* **15**, 2122–2133. doi: 10.1101/gad.204401

Sánchez PA, Cochrane TT (1980) Soil constraints in relation to major farming systems of tropical America. In 'Priorities of alleviating soil related constraints to food production in the tropics'. pp. 107–139. (IRRI: Los Baños)

Siegel S (1956) 'Nonparametric statistics: for the behavioral sciences.' (McGraw-Hill: New York)

Smith FW (2001) Sulphur and phosphorus transport systems in plants. *Plant and Soil* **232**, 109–118. doi: 10.1023/A:1010390120820

Uhde-Stone C, Zinn KE, Ramírez-Yañez M, Li A, Vance CP, Allan DL (2003) Nylon filter arrays reveal different gene expression in proteoid roots of white lupin in response to phosphorus deficiency. *Plant Physiology* **131**, 1064–1079. doi: 10.1104/pp.102.016881

Utriainen J, Holopainen T (2001) Influence of nitrogen and phosphorus availability and ozone stress on Norway spruce seedlings. *Tree Physiology* **7**, 447–456.

Manuscript received 26 April 2006, accepted 6 July 2006